ehp
ENVIRONMENTAL
HEALTH
PERSPECTIVES
http://www.ehponline.org

CpG Sites Associated with Cigarette Smoking: Analysis of Epigenome-Wide Data from the Sister Study

Sophia Harlid, Zongli Xu, Vijayalakshmi Panduri,
Dale P. Sandler, and Jack A. Taylor

http://dx.doi.org/10.1289/ehp.1307480
Received: 8 August 2013
Accepted: 2 April 2014
Advance Publication: 4 April 2014

NIH National Institute of
Environmental Health Sciences

CpG Sites Associated with Cigarette Smoking: Analysis of Epigenome-Wide Data from the Sister Study

Sophia Harlid,[1*] Zongli Xu,[2*] Vijayalakshmi Panduri,[1] Dale P. Sandler,[2] and Jack A. Taylor[1,2]

[1]Laboratory of Molecular Carcinogenesis, and [2]Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, North Carolina, USA; *these authors contributed equally to this work

Address correspondence to Jack A. Taylor, National Institute of Environmental Health Sciences, MD A3-05, 111 T.W. Alexander Drive, PO Box 12233, Research Triangle Park, NC 27709 USA. Telephone: (919) 541-4631. Fax : (919) 541-2511. E-mail: taylor@niehs.nih.gov

Running Title: Cigarette smoking and DNA methylation in blood

Acknowledgments: We thank the women who volunteered to participate in the Sister Study, the NIH Center for Inherited Disease Research, and the NIEHS Molecular Genetics Core and Microarray Core. This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01 ES044005, Z01 ES044032, and Z01 ES049033).

Competing Financial Interests: The authors declare no competing interests.

Abstract

Background: Smoking increases the risk of many diseases, and is also linked to blood DNA methylation changes that may be important in disease etiology.

Objectives: To identify novel CpG sites associated with cigarette smoking.

Methods: We used two epigenome-wide datasets from the Sister Study to identify and confirm CpG sites associated with smoking. One included 908 women with methylation measurements at 27,578 CpG sites using the Illumina HumanMethylation27 BeadChip; the other included 200 women with methylation measurements for 473,844 CpG sites using the Illumina HumanMethylation450 BeadChip. Significant CpGs from the second dataset that were not included in the 27k assay were validated by pyrosequencing in a subset of 476 samples from the first dataset.

Results: Our study successfully confirmed smoking associations for nine previously established CpGs and identified two potentially novel CpGs: cg26764244 in *GNG12* (p= 9.0×10^{-10}) and cg22335340 in *PTPN6* (p=2.9×10^{-05}). In addition to this we provide the first independent confirmation of the association between smoking status and cg02657160 in *CPOX* (p=7.28x10-^{7}). All 12 CpGs were undermethylated in current smokers and showed increasing percent methylation in former and never smokers.

Conclusions: We identified two potentially novel smoking related CpG sites, and provided independent replication of ten previously reported CpGs sites related to smoking, one of which is situated in the gene *CPOX*. The corresponding enzyme is involved in heme biosynthesis and smoking is known to increase heme production. Our study extends the evidence base for smoking-related changes in DNA methylation.

Introduction

Cigarette smoking is a known risk factor and contributor to adverse health outcomes such as cancer. The health effects associated with smoking are mediated through a variety of mechanisms including direct DNA damage, increased inflammation, and others (Hecht 2012; Mathers and Loncar 2006; Takahashi et al. 2010; Witschi et al. 1997). One of the ways by which smoking contributes to disease may be through epigenetic changes such as DNA methylation (Portela and Esteller 2010).

Epigenome wide association studies (EWAS) of DNA methylation in blood have recently identified a number of CpG loci associated with adult smoking (Breitling et al. 2011; Shenker et al. 2013a; Sun et al. 2013; Wan et al. 2012; Zeilinger et al. 2013) The first EWAS to investigate the association between smoking and DNA methylation made use of the 27K array and identified a differentially methylated CpG site in the gene *F2RL3* that showed decreased methylation in smokers (Breitling et al. 2011). More recent studies all made use of the newer 450K array bead chip (Shenker et al. 2013a; Sun et al. 2013; Zeilinger et al. 2013) to uncover new associations between smoking and DNA methylation and they have all confirmed the initial results in *F2RL3*. Another gene, *AHRR*, has consistently been a top finding in all of these studies.

One large EWAS investigated the association between DNA methylation in cord blood due to maternal smoking (Joubert et al. 2012). The study used the 450K methylation bead chip and the top hit here was also *AHRR*. There are also several reports of smoking-associated changes in global DNA methylation as well as changes in candidate genes (Breitling et al. 2012; Hillemacher et al. 2008; Kiseljak-Vassiliades and Xing 2011; Philibert et al. 2010; Smith et al. 2007; Tekpli et al. 2012; Wolff et al. 2008; Zhang et al. 2011).

Two of the five EWAS (Breitling et al. 2011; Wan et al. 2012) made use of the 27K methylation array from Illumina which assesses methylation at 27,578 CpG sites, however this array lacks coverage at many potentially relevant sites. Three studies made use of the newer 450K methylation array which measures methylation at 473,844 CpG sites and only one of these had a substantial sample size (Zeilinger et al. 2013).

We examined smoking and DNA methylation in a nationwide sample of volunteer women to confirm previously reported results as well as identify new sites associated with smoking. We analyzed two sets of methylation array data from the Sister Study- a prospective cohort study focused on environmental and familial risk factors for breast cancer and other diseases in women with a sister diagnosed with the disease.

Methods

Study design

All study subjects were participants of the NIEHS Sister Study, a nationwide prospective cohort study designed to examine genetic and environmental determinants of breast cancer. To be eligible for the Sister Study, women could not have had breast cancer themselves but must have had a biological sister with breast cancer. Detailed information can be found at http://sisterstudy.niehs.nih.gov (NIEHS 2014). At baseline all participants provided a blood sample and completed an extensive questionnaire on smoking history, including current smoking status, years since quitting, amount, and duration. Information on questionnaires is available through the Sister Study website cited above. Informed consent was obtained from all participants prior to participation. The study was approved by the Institutional Review Boards of the National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health, and the Copernicus Group (http://www.cgirb.com/irb-services/).

Our first sample included 908 non-Hispanic white women between the ages of 35–75 years. Blood methylation status in these women had been determined previously using the HumanMethylation27 BeadChip as part of a case-cohort study of methylation and breast cancer. The original study (Xu et al. 2013) utilized a nested case-cohort study design and included all incident breast cancer cases diagnosed between blood draw and May 2008 (n=329). As controls a random sample of women was drawn from the 29,026 participants enrolled in the Sister Study by June 2007 (n=709). Some samples were later excluded due to lack of DNA or poor quality measurements. For our study we only included non-Hispanic white women to account for factors affected by race. Our final sample therefore included 296 women who had been diagnosed with cancer within 46 months of blood draw and a random sample of 612 women who had remained cancer free for up to 55 months of follow up (Xu et al. 2013). We refer to this group as the 27K dataset.

Our second group of Sister Study participants included 200 women whose blood methylation was measured as part of an unpublished pilot study examining potential prenatal exposure to diethylstilbestrol (DES). In this study all women were non-Hispanic white between the ages of 41 and 59 years. All women in this sample set were originally selected from a subset of 1,802 women from a special sub-study validating self-reported exposures with the participant's mothers. A total of 100 exposed and 100 unexposed non-Hispanic white women were selected (unexposed women were frequency matched on age). Samples from all 200 women were successfully assayed using the HumanMethylation450 BeadChip. We refer to this group as the 450K dataset.

Infinium methylation assays

The Illumina HumanMethylation27 BeadChip provides DNA methylation data at single CpG site resolution for 27,578 different CpG sites covering promoter regions for over 14,000 Human RefSeq genes across 23 chromosomes. The Illumina HumanMethylation450 BeadChip provides information on 485,577 CpG sites with coverage of 99% of RefSeq genes and an average of 17 CpG sites per gene including sites in the promoter, 5'UTR, first exon, gene body, and 3'UTR. 25,978 CpG sites are shared between HumanMethylation27 and HumanMethylation450 BeadChip.

DNA from both sets of women was extracted from frozen whole blood samples as previously described (Xu et al. 2013). Extracted DNA was quantified using Quant-iT^{TM} PicoGreen dsDNA reagent (Invitrogen, Carlsbad CA) and stored at -20°C. One microgram of DNA was bisulfite converted using the EZ-DNA Methylation kit (Zymo Research, Irvine CA) following the manufacturer’s protocol. DNA was hybridized to the Illumina Human Methylation27 or Illumina HumanMethylation450 BeadChip arrays (Illumina, San Diego CA) following the manufacturer’s protocol and then scanned with an Illumina iScan (Illumina, San Diego CA). Data was analyzed using Illumina GenomeStudio® software (version 2011.1).

At each CpG site on the array, methylation status was determined based on intensity measures of two probes corresponding to unmethylated (U) or methylated (M) CpGs. Dye bias between U and M for type II probes in HumanMethylation450 BeadChip was corrected using the normalizeMethyLumiSet method in R package “methylumi”. Before association analysis, the intensity values were separately Robust Multichip Average (RMA) background corrected (Irizarry et al. 2003) and quantile normalized across arrays. The methylation level (beta value) of a specific CpG site was calculated as the ratio of normalized fluorescent intensities between methylated and unmethylated alleles $\beta=M/(M+U+100)$. To avoid SNP effects on methylation measures, we excluded CpG probes with SNPs present at target sites (428 CpGs from HumanMethylation27 and 20869 CpGs from HumanMethylation450 BeadChip). In both datasets a non-specific filtering step was applied to filter out the 20% CpGs with the smallest inter-quartile-range (IQR) of methylation values before association analysis. In the 27K dataset we tested 21,659 probes and in the 450K dataset we tested 369,120 probes.

Pyrosequencing

Pyrosequencing assays were developed for two CpGs (cg02657160 in *CPOX* and cg15999356 in *YAP1*) that showed potentially novel associations with smoking in the 450K dataset. Neither of the CpGs was present on the HumanMethylation27 BeadChip and thus the women who were members of the 27K dataset had not been evaluated at these two sites. For pyrosequencing analysis we selected 68 current smokers from the 27K dataset, and for each current smoker we selected three former smokers and three never smokers matched for age at blood draw and bisulfite conversion batch (Table 1).

Pyrosequencing primers for cg02657160 (*CPOX*) and cg15999356 (*YAP1*) were designed using Pyromark Assay Design version 2.0.2.15 (Qiagen, Valencia CA) (Supplemental Material, Table S1). Reaction mixtures (25 μL) containing 100ng of bisulfite converted DNA, 5pmol of each primer (forward and reverse) PCR buffer (Invitrogen, Carlsbad CA), 3mM $MgCl_2$, 1mM dNTP, and 0.8 units of taq polymerase (Invitrogen, Carlsbad CA) were heated to 95°C for 15 minutes, followed by 45 PCR cycles (95°C for 20 seconds, 55°C for 20 seconds and 72°C for 20 seconds) with a final extension at 72^{o}C for 5 minutes. Following PCR, the biotin labeled PCR product was hybridized to streptavidin-coated sepharose beads (GE Healthcare, Madison WI), and denatured in 0.2 M NaOH to provide a single-stranded sequencing template. Pyrosequencing primers (0.3 μmol/L) were annealed to the single-stranded template and the pyrosequencing was carried out using PyroMark Q96 MD System (Qiagen, Valencia CA) according to the manufacturer's instructions. Percentage methylation was quantified using the Pyro Q-CpG Software (Qiagen, Valencia CA).

Statistical analysis

A multiple linear regression model was employed to test the association between smoking status (never, former, and current smoker - see Smoking Variable, Supplemental Material for a complete description of this variable) and methylation (β value) at each CpG site adjusting for age at blood draw. In the 27K dataset we also adjusted for case-control status and in the 450K dataset we adjusted for DES exposure. A surrogate variable analysis (Leek and Storey 2007) was carried out in the methylation β value matrix to derive an additional set of variables. We then adjusted for these variables in the multiple linear regression model, hereby accounting for known and unknown confounders such as experimental batch effect. To examine whether additional variables such as passive smoking were influencing our results, we analyzed the correlation between our primary smoking variables and passive smoking, maternal smoking, parity, and menopausal status as well by adjusting for these variables. All analyses were repeated in the breast cancer free women only (27K set) but not in the DES-unexposed women from the 450K dataset as this would be a highly underpowered analysis due to the small sample size.

To minimize the effect of outliers, for each CpG probe we excluded approximately 0.8% of the β values that were more than three standard deviations from the mean. To correct for multiple testing, we estimated the false discovery rate (FDR) using the q-value framework (Storey and Tibshirani 2003). Effects of years of smoking, pack-year, and time since quitting on methylation changes were tested using multiple linear regression adjusting for age at blood draw. When analyzing the pyrosequencing data we employed a two-way ANOVA method to model the 1:3:3 matched study design and tested whether the methylation level in current smokers or former smokers differed from the methylation level never smokers.

Results

Characteristics of study participants according to smoking status are summarized in Table 1. Former smokers were slightly older than never smokers and current smokers. Current smokers had, on average, longer smoking history than former smokers. The smoking characteristics of women in the 27K dataset did not differ between breast cancer cases and non-cases; we therefore included both groups in the analysis with adjustment for case-status. This was also true for the 450K dataset where there was no difference between exposed and unexposed women. The 27K dataset included 496 (54.6%) never smokers, 344 (37.9%) former smokers, and 68 (7.5%) current smokers and the 450K dataset included 118 (59%) never smokers, 70 (35%) former smokers and 12 (6%) current smokers. The subset of the 27K dataset that was used for pyrosequencing included 204 (42.8%) never smokers, 204 (42.8%) former smokers and 68 (14.2%) current smokers (Table 1).

Using the 27K dataset as a discovery set, there were 18 CpGs associated with smoking status at false discovery rate of $q<0.05$ (Figure 1A, Figure 2A, and Supplemental Material, Table S2).

Seventeen of the 18 identified CpGs were available for replication analysis in the 450K dataset (cg13185177 was not present on the 450K array and thus could not be further examined). Eight of the 17 CpGs were confirmed in the 450K dataset at an *a priori* p-value threshold of 0.05 (Table 2). Although the percent methylation estimates for each CpG varied somewhat between the two array platforms, differences between never smokers, former, and current smokers for each of the 8 CpGs were similar between the two array platforms (Table 2). Smoking associations had been reported previously for six of the eight CpGs, while associations for the remaining two (cg22335340 in *PTPN6* and cg26764244 in *GNG12*) had not been previously reported (Table 2).

When analyzing the correlation between the primary smoking variables and other potential confounders, second hand smoking exposure was highly correlated with the primary smoking variables whereas none of the other potential confounders were correlated (data not shown). After adjustment, all except for one of the probes (cg22335340 in PTPN6) reported in Table 2 were still statistically significant at FDR threshold of 0.05; there was no statistically significant association with passive smoking (Supplemental Material, Table S3). We repeated the analysis in the 612 breast cancer free women from the 27K dataset (Supplemental Material, Table S4). When repeating the analysis in NonCases, six of the probes remained significant and two probes (cg13500388 in CBFB and cg11314684 in AKT3) failed to reach significance compared to when we used the dataset containing both Cases + NonCases. This could possibly be due to small sample size.

We also used the 450K dataset as a second discovery set to look for new CpGs associated with smoking. Because of the smaller sample size in this dataset we combined former smokers and current smokers into a single group for smoking association analysis. After excluding the 17 CpGs already examined as part of the replication study, five CpGs were significantly associated with smoking status at a study-wide FDR threshold of 0.05 (Figure 1B, Figure 2B, Table 3). Three of the five CpGs had previously been reported to be associated with smoking while two, cg15999356 in *YAP1* and cg02657160 in *CPOX,* had not been previously reported at the time our experiments were designed.

As an independent replication we used pyrosequencing to examine the *YAP1* and *CPOX* CpGs in 476 women who originally participated in the 27K study (which did not include these two sites on the 27K array). Although cg15999356 in *YAP1* showed little evidence of association with smoking, (P=0.23), there was strong evidence of association for cg02657160 in *CPOX* (Table 3 and Supplemental Material, Figure S1, two-way ANOVA test $P=1.1\times10^{-11}$).

Including all CpGs discovered and replicated in the 27K dataset (eight CpGs, Table 2), and the 450K dataset (four CpGs, Table 3) our study could confirm smoking associations for ten established and two potentially novel CpGs. There was a consistent ordering of methylation values across all but two of the 12 CpGs with highest values in never smokers, intermediate values in former smokers, and lowest values in current smokers; all 12 CpGs exhibited the lowest methylation values in current smokers (Table 2 and Table 3).

Among former smokers methylation values for all 12 CpG sites showed decreasing methylation with increasing number of years smoked and increasing pack years, although these findings were not statistically significant at all sites (Supplemental Material, Tables S5 and S6). Similarly, among former smokers methylation levels consistently increased with time since quitting smoking, although again these results were not statistically significant at all sites (Supplemental

Material, Tables S5 and S6). These findings were somewhat less consistent (both for direction and statistical significance) among current smokers, perhaps reflecting the smaller sample size of this group. Using multiple linear regression we found significant interaction between the number of years smoked and time since quitting on methylation for three CpG sites: cg03636183 at *F2RL3* (P=2.73×10^{-4}), cg19859270 at GPR15 (P=1.98×10-^{2}), and cg22335340 at *PTPN6* (P=2.91×10^{-3}). These three CpGs were also inversely associated with smoking pack years, but the association disappeared after adjusting for number of years smoked (data not shown).

Discussion

We performed an EWAS to investigate the effect of cigarette smoking on DNA methylation using previously collected data from the Sister Study. We confirmed (at a threshold of $p<0.05$) smoking associated CpGs using pyrosequencing for nine CpGs (eight discovered in the 27K dataset and 1 discovered in the 450K dataset). Two of these nine CpGs remain unreported (cg22335340 in *PTPN6* and cg26764244 in *GNG12*). The third CpG (cg02657160 in *CPOX*), was recently reported as a supplementary result (Zeilinger et al. 2013) but aside from this its association with smoking has not been described. We also summarize reported smoking-associated CpGs and our association p values in Supplemental Material, Table S7.

The most noteworthy finding of our study is the fact that we were able to successfully replicate a large number of previously identified CpG sited associated with cigarette smoking exposure. This further strengthens the validity of their association with smoking and contributes information about the magnitude of effect that smoking has on DNA methylation levels in blood.

In line with most previous studies we found that DNA methylation varies between current smokers and former smokers for most significant CpG sites. This may make it a useful biomarker of smoking status (Shenker et al. 2013b) and also indicates that these changes are stable for years and may continue to contribute to the increased risk of adverse outcomes associated with smoking.

The confirmation of cg19859270 in CPOX is particularly interesting as it is situated only 60kbp upstream of another smoking-associated CpG in GPR15 (cg19859270) that was also significant in our study and was previously identified by Wan et al 2012 (Figure 3). *GPR15* codes for a g-protein coupled receptor and functions as a co-receptor for human immunodeficiency virus (Blaak et al. 2005). Interestingly, gene expression of *GPR15* has been shown to increase in B cells from smokers compared to never smokers (Pan et al. 2010). There are >450 CpG sites that lay between the *CPOX* CpG and the *GRP15* CpG but these sites are not included in either the 27K or 450K arrays and thus this region may be of interest for future study.

CPOX (coproporphyrinogen oxidase) converts coprotoporphyrin III to protoporphyriogen IX, a reaction that constitutes the sixth step in heme biosynthesis (Layer et al. 2010). Heme is central in the synthesis of hemoglobin and is also important for the function of a large number of other proteins including the cytochromes P450s (Correia et al. 2011). Rare inherited mutations in *CPOX* cause the autosomal dominant disease coproporphyria which, like other porphyrias, is associated with low hemoglobin levels and anemia (Puy et al. 2010; Schmitt et al. 2005). Smoking is known to increase hemoglobin demands and can trigger coproporphyria attacks (Layer et al. 2010; Lip et al. 1991; Sassa 2006) it is also associated with both higher red blood cell counts and increased metabolism of cytochrome P450 enzymes (Schwartz and Weiss 1994; Van Tiel et al. 2002). A plausible connection between smoking and *CPOX* is that smoking exposure increases demand for heme synthesis which may result in increased expression of

CPOX. The observed DNA methylation changes in CPOX, although small (≤ 2%), reflects a change in methylation, and presumably gene expression, in a small population of cells, although average gene expression across the cell population is unlikely to be altered.

We also observed smoking associations with DNA methylation at two potentially novel CpG sites: cg22335340 in *PTPN6* and cg26764244 in *GNG12*. The *PTPN6* CpG is located 500bp upstream of the transcriptional start site. *PTPN6* codes for the protein SHP-1, a protein tyrosine phosphatase that is a putative tumor suppressor gene (Wu et al. 2003). It is expressed in hematopoietic cells and expression in lymphocytes from smokers is increased relative to levels among nonsmokers (Charlesworth et al. 2010), suggesting perhaps that the decreased DNA methylation we observe in smokers may be associated with increased transcriptional activity.

Cg26764244 in *GNG12* is also previously unreported in relation to smoking. However, two studies report an association between smoking and DNA methylation in another CpG in the same gene (cg25189904) (Shenker et al. 2013a; Zeilinger et al. 2013). Both cg26764244 and the previously reported cg25189904 are part of a close group of five CpGs (cg03140521, cg13184736, cg13399816, cg25189904 and cg26764244) situated in the south shore of a CpG island spanning the *GNG12* promoter (350bp from the transcriptional start site). In our 450K data set all five CpGs show decreased methylation in smokers and all have p-values <0.05. The CpG reported by Shenker et al. and Zeiliger et al. (cg25189904) is not included on the 27K array but when we examine the methylation values in the 450K dataset it shows the largest methylation decrease and the smallest p-value of the CpGs in the five CpG cluster ($p=2.94x10^{-6}$) suggesting that cg25189904 may be a CpG of interest.

We were also able to positively confirm (at genome wide significance) eight CpGs in seven loci that have previously been described by others (*F2RL3, LRRN3, AKT3, CNTNAP2, CFBF, AHRR,* and 2q37.1) (Breitling et al. 2011; Shenker et al. 2013a; Sun et al. 2013; Wan et al. 2012). The top smoking-related finding in several studies is represented by cg05575921 in *AHRR*. This gene codes for the aryl hydrocarbon receptor repressor and is a component of the AhR signaling cascade where it functions as a pathway inhibitor. The transcription factor AhR triggers expression of a diverse set of genes, some of which are involved in metabolism of endogenous substances including toxins from cigarette smoke (Haarmann-Stemmann et al. 2007; Kawajiri and Fujii-Kuriyama 2007; Swedenborg and Pongratz 2010). Additionally, the relationship between smoking and the *AHRR* CpG has been identified in studies using DNA from both peripheral blood and pulmonary macrophages (Monick et al. 2012).

Similar to findings from previous reports (Wan et al. 2012; Zeilinger et al. 2013; Shenker et al. 2013b) we find evidence of an inverse association between pack-years and DNA methylation and a positive association between time since quitting smoking and DNA methylation. However, in addition to pack years we also investigated years of smoking and found this inverse association to be even stronger than that between packs years and DNA methylation.

The main strengths of this study are the large sample size (>1100 women), confirmation of our 27K results in DNA from an independent group of women run on the 450K array and verification by pyrosequencing. Limitations include few current smokers in both the 27K and the 450K dataset and use of self-report for smoking information which is usually reliable although sometimes underreported (Wagenknecht et al. 1992). Another possible limitation is that we have chosen to combine former smokers with current smokers when analyzing the 450K dataset.

There were only 12 current smokers in that dataset, so that estimates for current smokers would be unstable. Previous studies (Zeilinger et al. 2013; Shenker et al. 2013a; Shenker et al. 2013b) indicate that former smokers might be more similar to never smokers than to current smokers but this would only bias our results towards the null and is therefore unlikely to result in false positive associations.

Conclusions

In this study we report 12 CpGs that are differentially methylated in smokers compared to nonsmokers, two of which had not been previously described. Most notably we provide the first independent replication of a cg02657160 in *CPOX,* a gene responsible for heme synthesis. We report decreased DNA methylation in smokers at this CpG*,* and suggest that epigenetic changes at this locus may reflect smoking-related demands for heme biosynthesis.

References

Blaak H, Boers P, Gruters R, Schuitemaker H, Van Der Ende M, Osterhaus A. 2005. CCR5, GPR15, and CXCR6 are major coreceptors of human immunodeficiency virus type 2 variants isolated from individuals with and without plasma viremia. J Virol 79:1686-1700.

Breitling LP, Salzmann K, Rothenbacher D, Burwinkel B, Brenner H. 2012. Smoking, F2RL3 methylation, and prognosis in stable coronary heart disease. Eur Heart J 33:2841-2848.

Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. 2011. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. Am J Hum Genet 88:450-457.

Charlesworth J, Curran J, Johnson M, Göring H, Dyer T, Diego V, et al. 2010. Transcriptomic epidemiology of smoking: the effect of smoking on gene expression in lymphocytes. BMC Med Genomics 3:1-11.

Correia M, Sinclair P, De Matteis F. 2011. Cytochrome P450 regulation: the interplay between its heme and apoprotein moieties in synthesis, assembly, repair, and disposal. Drug Metab Rev 43:1-26.

Haarmann-Stemmann T, Bothe H, Kohli A, Sydlik U, Abel J, Fritsche E. 2007. Analysis of the transcriptional regulation and molecular function of the aryl hydrocarbon receptor repressor in human cell lines. Drug Metab Dispos 35:2262-2269.

Hecht S. 2012. Lung carcinogenesis by tobacco smoke. Int J Cancer 131:2724-2732.

Hillemacher T, Frieling H, Moskau S, Muschler MA, Semmler A, Kornhuber J, et al. 2008. Global DNA methylation is influenced by smoking behaviour. Eur Neuropsychopharmacol 18:295-298.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4:249-264.

Joubert BR, Haberg SE, Nilsen RM, Wang X, Vollset SE, Murphy SK, et al. 2012. 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. Environ Health Perspect 120:1425-1431.

Kawajiri K, Fujii-Kuriyama Y. 2007. Cytochrome P450 gene regulation and physiological functions mediated by the aryl hydrocarbon receptor. Arch Biochem Biophys 464:207-212.

Kiseljak-Vassiliades K, Xing M. 2011. Association of Cigarette Smoking with Aberrant Methylation of the Tumor Suppressor Gene RARbeta2 in Papillary Thyroid Cancer. Front Endocrinol (Lausanne) 2:1-5.

Layer G, Reichelt J, Jahn D, Heinz D. 2010. Structure and function of enzymes in heme biosynthesis. Protein Sci 19:1137-1161.

Leek J, Storey J. 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet 3:1724-1735.

Lip G, McColl K, Goldberg A, Moore M. 1991. Smoking and recurrent attacks of acute intermittent porphyria. BMJ 302:507.

Mathers C, Loncar D. 2006. Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med 3:2011-2030.

Monick MM, Beach SR, Plume J, Sears R, Gerrard M, Brody GH, et al. 2012. Coordinated changes in AHRR methylation in lymphoblasts and pulmonary macrophages from smokers. Am J Med Genet B Neuropsychiatr Genet 159B:141-151.

NIEHS (National Institute of Environmental Health Sciences) 2014. The Sister Study Cohort Description. Available: http://sisterstudy.niehs.nih.gov [accessed 19 March 2014].

Pan F, Yang T-L, Chen X-D, Chen Y, Gao G, Liu Y-Z, et al. 2010. Impact of female cigarette smoking on circulating B cells in vivo: the suppressed ICOSLG, TCF3, and VCAM1 gene functional network may inhibit normal cell function. Immunogenetics 62:237-251.

Philibert RA, Beach SR, Gunter TD, Brody GH, Madan A, Gerrard M. 2010. The effect of smoking on MAOA promoter methylation in DNA prepared from lymphoblasts and whole blood. Am J Med Genet B Neuropsychiatr Genet 153B:619-628.

Portela A, Esteller M. 2010. Epigenetic modifications and human disease. Nat Biotechnol 28:1057-1068.

Puy H, Gouya L, Deybach J-C. 2010. Porphyrias. Lancet 375:924-937.

Sassa S. 2006. Modern diagnosis and management of the porphyrias. Br J Haematol 135:281-292.

Schmitt C, Gouya L, Malonova E, Lamoril J, Camadro J-M, Flamme M, et al. 2005. Mutations in human CPO gene predict clinical expression of either hepatic hereditary coproporphyria or erythropoietic harderoporphyria. Hum Mol Genet 14:3089-3098.

Schwartz J, Weiss S. 1994. Cigarette smoking and peripheral blood leukocyte differentials. Ann Epidemiol 4:236-242.

Shenker NS, Polidoro S, van Veldhoven K, Sacerdote C, Ricceri F, Birrell MA, et al. 2013a. Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. Hum Mol Genet 22: 843-851.

Shenker NS, Ueland PM, Polidoro S, van Veldhoven K, Ricceri F, Brown R, Flanagan JM, Vineis P. 2013b. DNA methylation as a long-term biomarker of exposure to tobacco smoke. Epidemiology 24:712-716.

Smith IM, Mydlarz WK, Mithani SK, Califano JA. 2007. DNA global hypomethylation in squamous cell head and neck cancer associated with smoking, alcohol consumption and stage. Int J Cancer 121:1724-1728.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100:9440-9445.

Sun Y, Smith A, Conneely K, Chang Q, Li W, Lazarus A, et al. 2013. Epigenomic association analysis identifies smoking-related DNA methylation sites in African Americans. Hum Genet 132:1027-1037

Swedenborg E, Pongratz I. 2010. AhR and ARNT modulate ER signaling. Toxicology 268:132-138.

Takahashi H, Ogata H, Nishigaki R, Broide D, Karin M. 2010. Tobacco smoke promotes lung tumorigenesis by triggering IKKbeta- and JNK1-dependent inflammation. Cancer cell 17:89-97.

Tekpli X, Zienolddiny S, Skaug V, Stangeland L, Haugen A, Mollerup S. 2012. DNA methylation of the CYP1A1 enhancer is associated with smoking-induced genetic alterations in human lung. Int J Cancer 131:1509-1516.

Van Tiel E, Peeters P, Smit H, Nagelkerke N, Van Loon A, Grobbee D, et al. 2002. Quitting smoking may restore hematological characteristics within five years. Ann Epidemiol 12:378-388.

Wagenknecht L, Burke G, Perkins L, Haley N, Friedman G. 1992. Misclassification of smoking status in the CARDIA study: a comparison of self-report with serum cotinine levels. Am J Public Health 82:33-36.

Wan ES, Qiu W, Baccarelli A, Carey VJ, Bacherman H, Rennard SI, et al. 2012. Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. Hum Mol Genet 21:3073-3082.

Witschi H, Espiritu I, Peake J, Wu K, Maronpot R, Pinkerton K. 1997. The carcinogenicity of environmental tobacco smoke. Carcinogenesis 18:575-586.

Wolff EM, Liang G, Cortez CC, Tsai YC, Castelao JE, Cortessis VK, et al. 2008. RUNX3 methylation reveals that bladder tumors are older in patients with a history of smoking. Cancer Res 68:6208-6214.

Wu C, Sun M, Liu L, Zhou G. 2003. The function of the protein tyrosine phosphatase SHP-1 in cancer. Gene 306:1-12.

Xu Z, Bolick S, Deroo L, Weinberg C, Sandler D, Taylor J. 2013. Epigenome-wide Association Study of Breast Cancer Using Prospectively Collected Sister Study Samples. J Natl Cancer Inst 105:694-700.

Zeilinger S, Kühnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, et al. 2013. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PloS one 8:1-14

Zhang FF, Cardarelli R, Carroll J, Fulda KG, Kaur M, Gonzalez K, et al. 2011. Significant differences in global genomic DNA methylation by gender and race/ethnicity in peripheral blood. Epigenetics 6:623-629.

Table 1. Sample characteristics [number(s) or mean ± SD].

Characteristic	Never smoker	Former smoker	Current smoker
27K dataset			
N (number)	496	344	68
Age (mean ± SD)	55.13(9.09)	57.07 (8.83)	53.18 (9.63)
Pack year (mean ± SD)	-	13.24 (13.93)	25.17 (13.58)
Years of smoking (mean ± SD)	-	15.13 (10.64)	32.64 (11.46)
Age started smoking (mean ± SD)	-	17.92 (3.96)	17.25 (3.57)
Years since quitting (mean ± SD)	-	22.03 (10.71)	-
Breast cancer case/ non-case (number)	165/331	111/233	20/48
450K dataset			
N (number)	118	70	12
Age (mean ± SD)	50.31(4.80)	51.16 (5.04)	50.50 (3.55)
Pack year (mean ± SD)		9.92 (11.46)	24.45 (15.34)
Years of smoking (mean ± SD)	-	11.70 (9.06)	29.63 (6.84)
Age started smoking (mean ± SD)	-	17.41 (3.70)	16.25 (2.56)
Years since quitting (mean ± SD)	-	20.94 (9.35)	-
DES exposed/non-exposed (number)	60/58	34/36	6/6
Pyrosequencing			
N (number)	204	204	68
Age (mean ± SD)	53.14 (9.50)	56.11 (8.91)	53.18 (9.63)
Pack year (mean ± SD)		18.37 (14.71)	25.17 (13.58)
Years of smoking (mean ± SD)	-	20.44 (10.06)	32.64 (11.46)
Age started smoking (mean ± SD)	-	17.03 (3.22)	17.25 (3.57)
Years since quitting (mean ± SD)	-	16.79 (8.42)	-

Table 2. Differentially methylated CpG sites by smoking status at FDR threshold of 0.05 (27K dataset) and validated with a p<0.05 (450K dataset).

CpG	Gene	27K: β-value never	27K: β-value former	27K: β-value current	27K: P	27K: FDR	450K: β-value never	450K: β-value former	450K: β-value current	450K: P	Source[a]
cg03636183	F2RL3	0.85	0.83	0.78	9.3×10^{-70}	1.9×10^{-65}	0.79	0.77	0.7	2.0×10^{-07}	(Breitling et al. 2011; Shenker et al. 2013a; Sun et al. 2013; Wan et al. 2012; Zeilinger et al 2013)
cg19859270	GPR15	0.84	0.83	0.81	2.9×10^{-53}	3.0×10^{-49}	0.95	0.94	0.93	3.8×10^{-04}	(Breitling et al. 2011; Sun et al. 2013; Wan et al. 2012; Zeilinger et al 2013)
cg09837977	LRRN3	0.76	0.75	0.74	1.7×10^{-14}	1.2×10^{-10}	0.88	0.87	0.83	5.0×10^{-04}	(Sun et al. 2013; Wan et al. 2012)
cg26764244	GNG12	0.12	0.11	0.098	9.0×10^{-10}	4.6×10^{-06}	0.26	0.24	0.22	0.006	NA
cg16254309	CNTNAP2	0.063	0.06	0.053	1.9×10^{-09}	7.7×10^{-06}	0.064	0.06	0.039	0.01	(Shenker et al. 2013a; Sun et al. 2013; Wan et al. 2012; Joubert et al. 2012; Zeilinger et al 2013)
cg13500388	CBFB	0.43	0.42	0.41	1.8×10^{-07}	6.0×10^{-04}	0.55	0.54	0.51	0.02	(Sun et al. 2013)
cg11314684	AKT3	0.27	0.26	0.25	2.1×10^{-07}	6.2×10^{-04}	0.35	0.34	0.31	9.3×10^{-04}	(Sun et al. 2013)
cg22335340	PTPN6	0.65	0.65	0.64	2.9×10^{-05}	4.0×10^{-02}	0.78	0.77	0.75	0.01	NA

[a] Refers to previous publications where the CpG has been reported.

Table 3. Cigarette smoking related CpG sites identified in the 450K dataset and confirmation results by pyrosequencing experiment.

CpG	Gene	450K: β-value never	450K: β-value: former	450K: β-value: current	450K: β-value: P	450K: β-value: FDR	Pyro: β-value never	Pyro: β-value former	Pyro: β-value current	Pyro: P	Source[a]
cg02657160	CPOX	0.87	0.86	0.84	7.3×10^{-07}	4.5×10^{-02}	0.88	0.87	0.86	1.1x10^{-11}	(Zeilinger et al 2013)
cg15999356	YAP1	0.79	0.74	0.76	5.2×10^{-08}	1.9×10^{-02}	0.64	0.63	0.63	0.23	NA
cg05575921	AHRR	0.85	0.83	0.75	2.6×10^{-07}	2.4×10^{-02}	NA	NA	NA	NA	(Shenker et al. 2013a; Sun et al. 2013; Joubert et al. 2012; Zeilinger et al 2013)
cg06644428	2q37.1	0.1	0.081	0.043	2.6×10^{-07}	2.4×10^{-02}	NA	NA	NA	NA	(Shenker et al. 2013a; Sun et al. 2013; Zeilinger et al 2013)
cg05951221	2q37.1	0.56	0.53	0.41	7.1×10^{-07}	4.5×10^{-02}	NA	NA	NA	NA	(Shenker et al. 2013a; Sun et al. 2013; Zeilinger et al 2013)

[a] Refers to previous publications where the CpG has been reported.

Figure Legends

Figure 1. Manhattan plot of smoking epigenome wide association p-values. (A) Illumina HumanMethylation27 Beadchip, (B) HumanMethylaition450 Beadchip. Dotted horizontal line indicates false discovery rate of 0.05.

Figure 2. Q-Q plot for smoking epigenome-wide association p-values. Horizontal axis shows (-$\log_{10}$ transformed) expected p-values and the vertical axis shows (-$\log_{10}$ transformed) observed p-values. Panel A) Results from the 27K dataset. Panel B) Results from the 450K dataset.

Figure 3. Depiction of the region surrounding cg02657160. Cg02657160 is located in the first intron of *CPOX*, 60kbp from another CpG in *GPR15* that was found by both us and others (Wan et al 2012). This suggests that a larger area may be affected by methylation changes.

Figure 1.

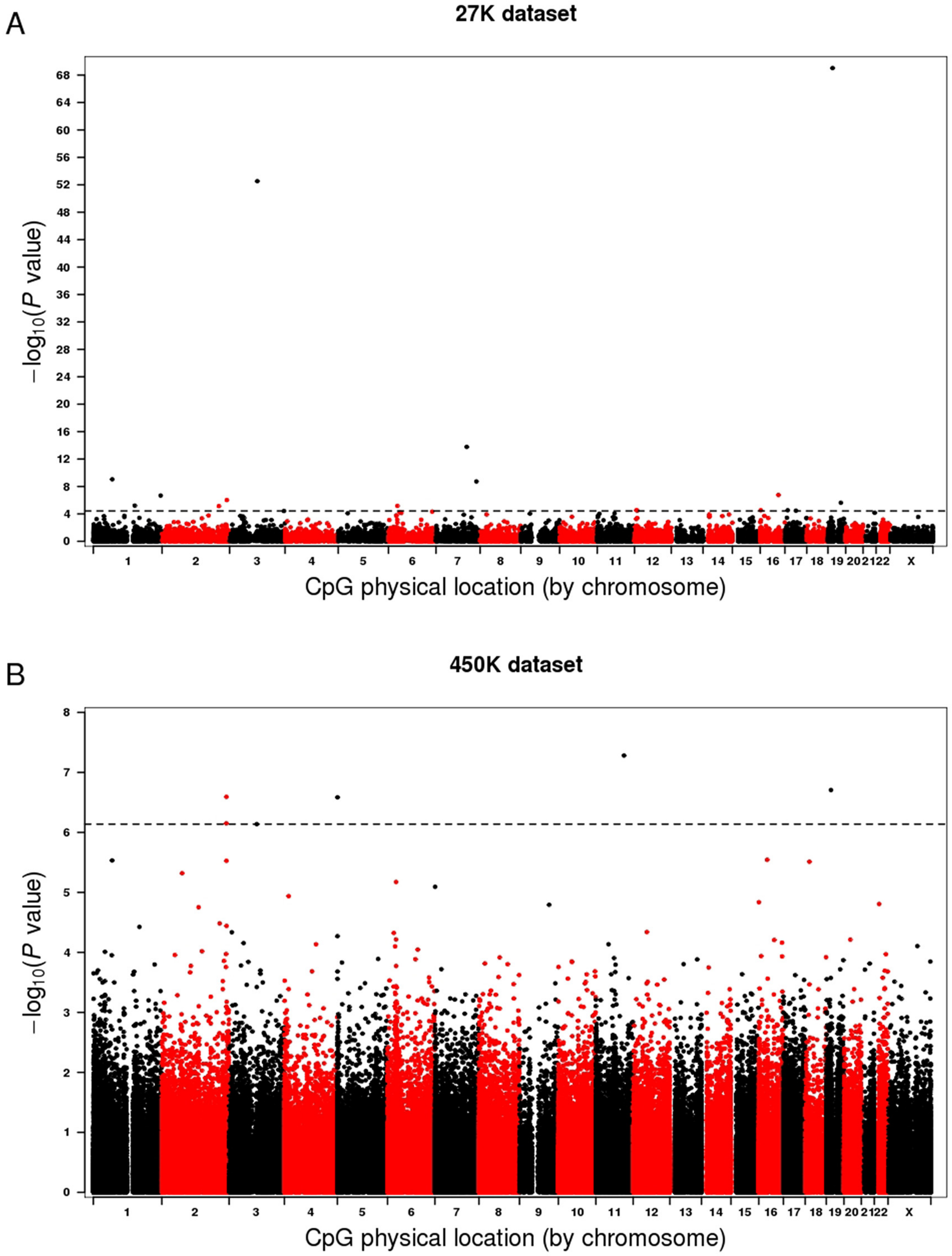

A
27K dataset
−log10(P value)
0
4
8
12
16
20
24
28
32
36
40
44
48
52
56
60
64
68
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
CpG physical location (by chromosome)
B
450K dataset
−log10(P value)
0
1
2
3
4
5
6
7
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
CpG physical location (by chromosome)

Figure 2.

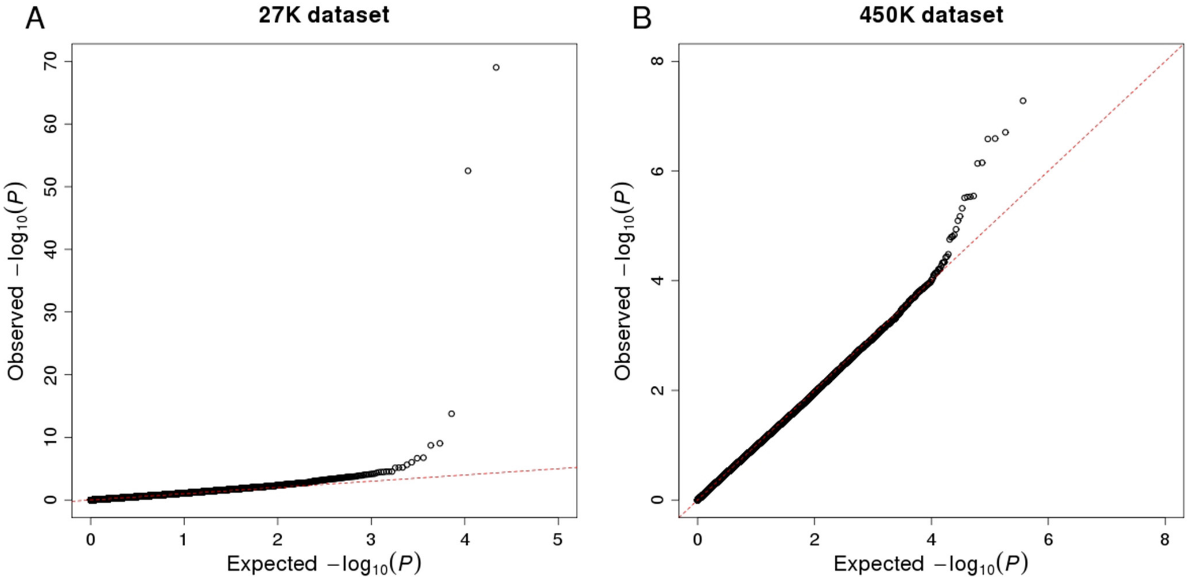
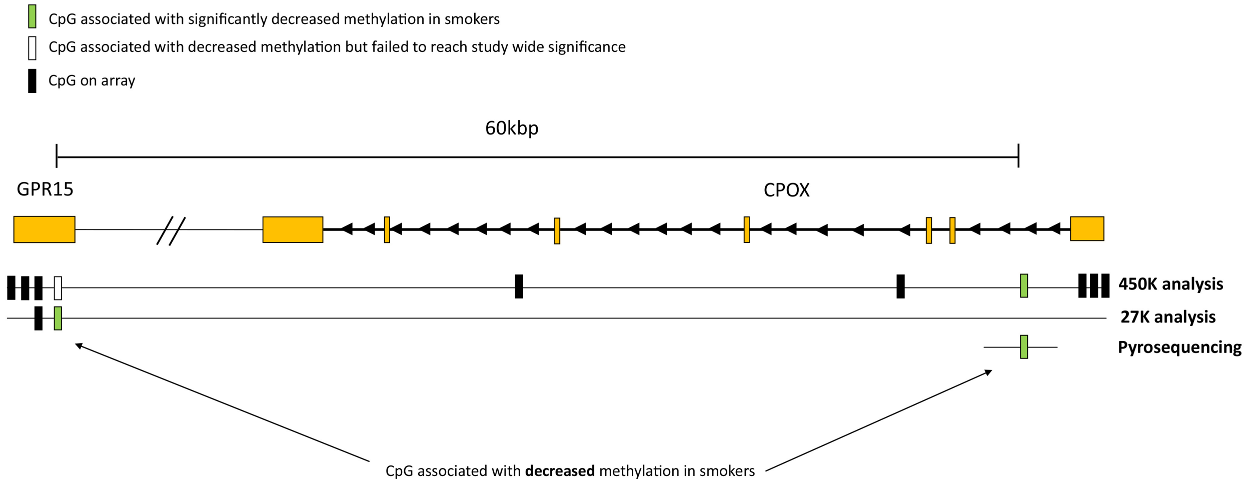

A
27K dataset
Observed $-\log_{10}(P)$
Expected $-\log_{10}(P)$
0
10
20
30
40
50
60
70
0
1
2
3
4
5
B
450K dataset
Observed $-\log_{10}(P)$
Expected $-\log_{10}(P)$
0
2
4
6
8
0
2
4
6
8

Figure 3.


CpG associated with significantly decreased methylation in smokers
CpG associated with decreased methylation but failed to reach study wide significance
CpG on array
60kbp
GPR15
CPOX
450K analysis
27K analysis
Pyrosequencing
CpG associated with **decreased** methylation in smokers